DOCUMENT RESUME

ED 399 293 TM 025 607

AUTHOR Johanson, George A.; Johanson, Susan N.
TITLE Differential Item Functioning in Survey Research.
PUB DATE Apr 96
NOTE 9p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; Data Analysis; *Evaluation Methods; Item Analysis; *Item Bias; *Research Methodology; *Surveys; *Test Construction; Test Items
IDENTIFIERS Item Bias Detection; *Mantel Haenszel Procedure

ABSTRACT

Differential item functioning (DIF), or item bias, occurs when individuals in a focal group respond differently to a test item than do individuals in a reference group even when comparisons are restricted to individuals with similar overall skill levels on the trait in question. It is common in constructing a questionnaire or survey to recommend that an item analysis be conducted in a manner similar to that used in cognitive measurement, but it is not common to be concerned with items as they perform differently. DIF has apparently not yet been widely recognized as a tool for developing a survey or for understanding survey responses. The Mantel Haenszel procedure is one of the empirical methods most commonly used to identify DIF, and its use in survey development is explored. Two examples, one involving the evaluation of student achievement for 777 male and 773 female elementary school students, and the other involving the evaluation of human service workshops for 798 participants aged over 40 years and 884 who were younger, illustrate the way in which information about DIF could have aided in the development of the instrument and interpretation of the data. DIF detection would seem a useful adjunct to the traditional item analysis that could be of substantial value at the pilot or revision stage of instrument development. (Contains 4 figures and 20 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
GEORGE JOHANSON
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Differential Item Functioning in Survey Research

George A. Johanson and Susan N. Johanson
College of Education, Ohio University

A paper presented at the annual meeting of the American Educational Research Association, New York, 1996

ED 399 293

Differential item functioning (DIF) or item bias occurs when individuals in a focal group respond differently to an item than individuals in a reference group even when the comparisons are restricted to individuals with similar overall skill levels on the trait in question. Typically, the total score on the instrument is used to stratify individuals in both focal and reference groups into a number of score-equivalent groups. If an item is more or less 'difficult' for one group than another when the comparisons are between subgroups with equal overall scores, then the item is said to function differentially for these groups. Of course, this procedure is only '...valid to the extent that the test as a whole is less biased than the individual items..' (Ironson, 1983).

DIF is not the same as *item impact* which occurs when two groups (reference and focal) differ in performance on an item (Dorans, 1989). Item impact is common and often expected. It is not newsworthy when fifth grade students outperform third grade students on a mathematics item. In contrast, DIF occurs when there is differential item performance between groups after conditioning on, or controlling for, overall skill level using a measure such as total score.

It is common when constructing a questionnaire or survey to recommend an item analysis be conducted in a similar manner to that used in cognitive measurement (e.g., Mueller, 1986). It is uncommon, however, to be concerned with identifying items that perform differentially. A search of the ERIC and PsychLit electronic databases using a combination of words and descriptors for attitude measurement and differential item functioning failed to turn up more than just a few entries (and most of these were not relevant). It would seem that DIF has not yet been widely recognized as a useful tool for developing a survey or for understanding survey responses even though the concept is certainly relevant to attitude assessment. DIF has obvious ethical implications but, even more generally, '...studies of measurement bias should be encouraged as part of the general process of construct validation...' (Millsap & Everson, 1993, p. 329). The purpose of this paper is to introduce and encourage the use of DIF analysis in survey research by illustrating its utility with two examples.

One possible reason for the seeming lack of interest in DIF by survey researchers might be that methods of detection generally assume a dichotomously scored (most often, right-wrong) item and many survey items admit more than two response categories. However, it may well be conceptually reasonable to recode graded response categories into two distinct (and exhaustive) categories such as 'agree/disagree', 'like/do not like', or 'good/not good'. If the resulting loss of information is unacceptable, there are more recent efforts afoot to detect DIF with polytomously scored (graded-response) items (Millsap & Everson, 1993; Cohen, Kim, & Baker, 1993; Welch & Hoover, 1993).

TM 025617

BEST COPY AVAILABLE

Interestingly, these efforts with graded-response items seem to be designed for use with cognitive (performance) items and not items assessing affect or opinion.

METHOD

Item bias can be investigated using either professional judgment and/or empirical methods. In survey research it is common practice to have items reviewed by knowledgeable individuals for a wide variety of errors of omission or commission (e.g., Converse & Presser, 1986). If there is anything we know from survey research it is that seemingly simple item rephrasing can have a large influence on responses. However, judgmental methods are likely not sufficient for the detection of DIF (Engelhard, Hansche, & Rutledge 1990), at least in cognitive assessment. This does not mean that the non-empirical review process is unnecessary or unsatisfactory, but only that it may be insufficient for analysis of DIF.

Our current discussion will be restricted to methods for binary items. Of the empirical methods for identifying DIF in binary items, the Mantel-Haenszel (MH) procedure (Mantel & Haenszel, 1959; Dorans, 1989) is often recommended (Holland & Thayer, 1988; Dorans & Holland, 1993). Other empirical methods for binary items include those based on item response theory, logistic regression, and standardization methods (Milsap & Everson, 1993). In the MH procedure, the overall measure (including the item in question) is used to form (K) groups of individuals with similar total scores. K is often simply the number of possible raw scores. For each level of K, form a 2x2 frequency table by crossing the item response (e.g., agree-disagree) with group membership (focus-reference). An overall odds-ratio is then computed from this Kx2x2 table as is a test statistic (approximately distributed as a χ^2 with one degree of freedom) testing the null that the odds-ratio equals one. This null is the hypothesis of no DIF. Using an agree-disagree interpretation of binary, the MH odds-ratio reflects the odds that an individual in the focal group will agree with the item under investigation when compared to an individual in the reference group (or vice-versa) when the individuals are matched on overall attitude. Particularly lucid presentations and explanations of the formulae can be found in Raju, Bode, & Larsen, 1989 or Camilli & Shepard, 1994.

Many standard statistical packages (e.g., SAS and BMDP) compute MH statistics since the procedure is often used in biostatistics (where the MH odds-ratio is frequently referred to as 'risk' or 'relative risk'). A macro for SPSS was recently made available (Nichols, 1994) and there are stand-alone programs (e.g., Fidalgo, 1994). The SPSS macro was used in this study. Recommendations regarding the number of subjects required for accurate use of the MH procedure call for a minimum of 100 persons in the smaller of the focal or reference groups and a total of 500 persons (Zieky, 1993). Uttaro and Millsap (1994) indicate that both the odds-ratio and the significance test are important in the detection of DIF. Of course, accumulated type I error rates can be a problem if many items are investigated. Raju, Bode, & Larsen (1989) suggest a Bonferroni-type correction be used.

It is usually recommended that the item under consideration be included in the total score used for stratification even though this would seem to possibly bias the process. In fact, Donoghue, Holland, & Thayer (1993) state that "We take the position that a proper application of the MH procedure to DIF studies must *include* the studied

item in the matching criterion..." (p. 152). The question of continuing to use items known to show DIF for subsequent analyses is less clear (e.g., see Clauser, Mazor, & Hambleton, 1993).

EXAMPLE ONE

The *Lead Teacher Project* (Martin, 1989) assessed elementary students skills in and attitudes towards mathematics and science. The science attitude assessment done in the Fall of 1992 with 1550 students in grades 1-6 is the current focus. The two forms (grades 1-3 and 4-6) of the survey used consisted of 10 common items (words and pictures describing science related activities) to which students responded that they enjoyed the activity (1) or did not enjoy the activity (0). There were 777 male students and 773 female students responding. Overall, the males (mean = .71) had a slightly more positive attitude towards science than the females (mean = .67). Gender defines our focal and reference groups for this example.

One item, #12 (one of the 10 common items), showed the picture of a dinosaur skeleton (a tyrannosaur with the caption 'Do you enjoy studying about living things from the past?') and was given relatively high ratings by both males (.84) and females (.71). The different ratings for this item from males and females may just be item impact. Figure 1, however, illustrates that the item functions differentially for males and females. For

<insert Figure 1 about here>

those familiar with item response theory (IRT), Figure 1 is actually two superimposed empirical item characteristic curves, one for males and the other for females. Many of the IRT methods of item bias detection use the area between item characteristic curves as a measure of DIF. Note that at virtually all attitude scores levels, males like #12 more than females. This is a good example of *uniform* or *consistent DIF* (Camilli & Shepard, 1994) in that the males prefer this item over females at all attitude levels. The MH odds-ratio for this item was .41 and can be interpreted to mean that the odds of a male liking this compared to a female is approximately 2.4:1 when the overall attitude towards science is the same. The odds-ratio difference from 1 is unlikely due to chance ($\chi^2 = 28.47$, $p < .0000$). In short, item #12 was biased in favor of males.

Since there were two forms of the survey, it was decided to look within each for evidence of DIF. With the grade 1-3 form (N = 600), the odds-ratio was .49, but the statistical test was non-significant ($\chi^2 = 2.67$, $p > .10$). Since difference in item preference is not uniform over attitude levels, this is referred to as *nonuniform* or *inconsistent DIF* (Camilli & Shepard, 1994). In the grade 4-6 form (955), the odds-ratio was .40 and was significantly different from 1 ($\chi^2 = 25.03$, $p < .0000$). Figures 2 and 3

<insert Figures 2 and 3 about here>

show the differences between forms or grade levels. While the item might possibly be retained for use in grades 1-3, it would certainly be omitted for grades 4-6. Usual item analyses showed that item #12 behaved well. The item-total correlation was .63 (.53 for the corrected item-total correlation) and so the item had relatively high discrimination.

Separate male and female item-total correlations were both above .60. The 10-item scale was quite reliable (coefficient alpha or KR-20 = .83). Without a DIF analysis, the item would likely remain in the scale where it would bias science attitudes towards males.

EXAMPLE TWO

The second example comes from an 8-item evaluation of human service workshops in Ohio in 1991. In this case, the focal and reference groups are those individuals less than 40 years of age (N = 884) and those 40 years of age or older (N = 798). The response categories for the items were POOR, FAIR, GOOD, EXCELLENT, and OUTSTANDING. The responses were recoded so that those answering GOOD or EXCELLENT were coded 1; all other responses were coded 0.

Item #6 asked whether the time allocated for the meeting was appropriate. The mean response to the recoded item was .71 and it correlated strongly (.72) with the 8-item total score (coefficient alpha = .88). While item #6 looks fine, there was some question as to whether it was biased towards the younger (and presumably less experienced) employee. That is, did the older employee tend to respond less favorably to this item than the younger employee even when the overall attitude of the individuals were the same? Figure 4 would indicate that this was not the case. The MH odds-ratio was computed to

<insert Figure 4 about here>

be 1.02 ($\chi^2 = .005$, $p > .94$). The item performed in a virtually identical fashion for the younger and older employees.

DISCUSSION

In brief, we have presented two examples with survey data where information about differential item functioning could have aided in both the development of the instrument and in the interpretation of the data. As an additional benefit, we feel that we have a better understanding of the responses of some groups to selected items after conducting the DIF analysis.

It was really quite surprising not to find applications of differential item functioning in the survey research literature. DIF detection would seem to be a useful adjunct to the traditional item analysis that could be of substantial value at the pilot or revision stage of instrument development. At a practical level, the MH procedure is available in several locations or forms and requires compromise only in that item responses must be in a binary format. Such item recoding seems a reasonable price to pay for so much potentially useful information.

REFERENCES

Camilli, G, & Shepard, L. A. (1994). Methods for identifying biased test items. Thousand Oaks, CA: Sage Publications.

Clauser, B., Mazor, K., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. Applied Measurement in Education, 6(4), 269-279.

Cohen, A. S., Kim, S. H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. Applied Psychological Measurement, 17(4), 335-350.

Converse, J. M., & Presser, S. (1986). Survey questions: Handcrafting the standardized questionnaire. Beverly Hills, CA: Sage Publications.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A monte-carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland, & H. Wainer (Eds.), Differential item functioning (pp. 137-166). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. Applied Measurement in Education, 2(3), 217-233.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland, & H. Wainer (Eds.), Differential item functioning (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Engelhard, G., Hansche, L., & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. Applied Measurement in Education, 3(4), 347-360.

Fidalgo, A. M. (1994). MHDIF: A computer program for detecting uniform and nonuniform differential item functioning with the Mantel-Haenszel procedure. Applied Psychological Measurement, 18(3), 300.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, & H. I. Braun (Eds.), Test validity (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ironson, G. H. (1983). Using item response theory to measure bias. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 155-174). Vancouver, B. C., Canada: Educational Research Institute of British Columbia.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Martin, R. (1989). The lead teacher project: K-6 mathematics and science elementary teacher enhancement. NSF grant number 91-47392.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. Applied Psychological Measurement, 17(4), 297-334.

Mueller, D. J. (1986). Measuring social attitudes: A handbook for researchers and practitioners. New York: Teachers College Press.

Nichols, D. P. (1994). The Mantel-Haenszel statistic for 2x2xK tables. Keywords: Tips and news for statistical software users, 54, 10-12.

Raju, N. S., Bode, R. K., & Larsen, V. S. (1989). An empirical assessment of the Mantel-Haenszel statistic for studying differential item performance. Applied Measurement in Education, 2(1), 1-13.

Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. Applied Psychological Measurement, 18(1), 15-25.

Welch, C., & Hoover, H. D. (1993). Procedures for extending item bias detection techniques to polytomously scored items. Applied Measurement in Education, 6(1), 1-19.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland, & H. Wainer (Eds.), Differential item functioning (pp. 337-347). Hillsdale, NJ: Lawrence Erlbaum Associates.

Figure 1. Proportion 'Liking' Item 12 by Gender at Different Score Levels for Students in Grades 1-6 (N=1550)

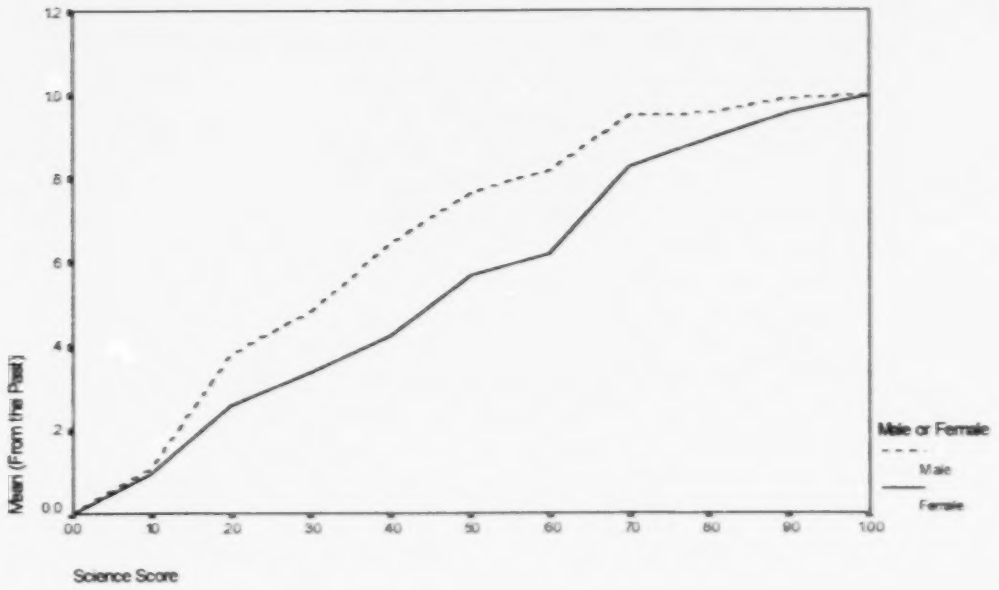

Figure 2. Proportion 'Liking' Item 12 by Gender at Different Score Levels for Students in Grades 1-3 (N=600)

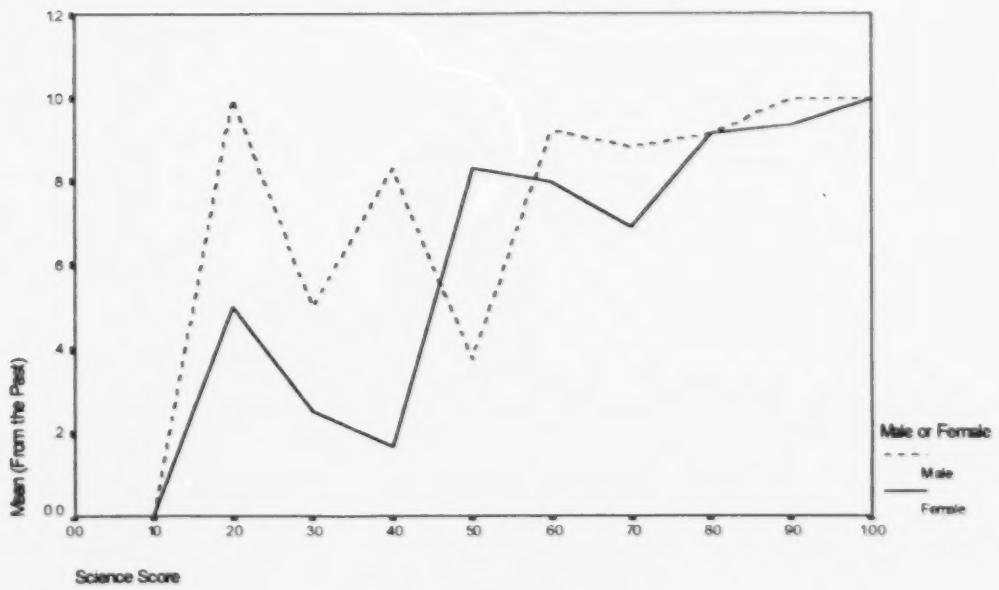

Figure 3. Proportion 'Liking' Item 12 by Gender at Different Score Levels for Students in Grades 4-6 (N=955)

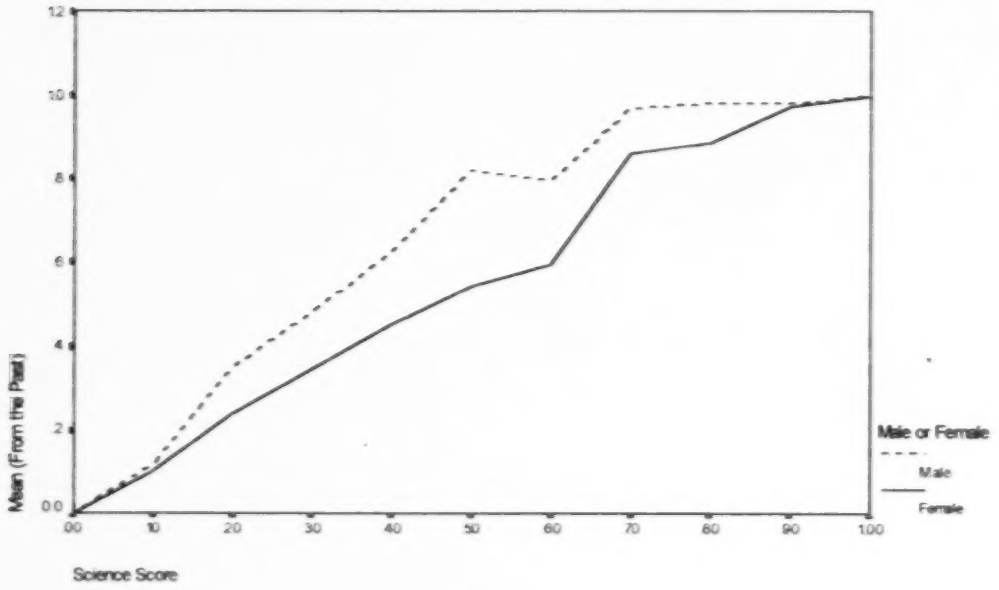

Figure 4. Proportion 'Excellent or Outstanding' Item 6 by Age Group at Different Score Levels for Participants (N=1351)